DOCUMENT RESUME

ED 218 361 TM 820 474

AUTHOR Livingston, Samuel A.; Kastrinos, William
TITLE A Study of the Reliability of Nedelsky's Method for Choosing a Passing Score.
INSTITUTION Educational Testing Service, Princeton, N.J. Center for Occupational and Professional Assessment.
REPORT NO ETS-RR-82-6
PUB DATE Mar 82
NOTE 18p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cutting Scores; *Judges; Multiple Choice Tests; Real Estate; *Scoring; Scoring Formulas; *Test Reliability
IDENTIFIERS Inter Rater Reliability; *Licensing Examinations; *Nedelsky Method

ABSTRACT

Leo Nedelsky developed a method for determining absolute grading standards for multiple choice tests. His method required a group of judges to examine each test question and eliminate those responses which the lowest D- student should be able to reject as incorrect. The correct answer probabilities remaining were used in computing an expected test score for the hypothetical test taker. The passing score was chosen to give the "F-D student" some probability of passing the test. Nedelsky's method of choosing passing scores are being used by many licensing and certification boards. Alternate forms of a test, containing 25 ittms in common, were used to measure the reliability of passing scores chosen by this method. Experts judged the tests. There were large and consistent differences between the judges for their passing scores. The experts tended to set higher passing scores the second time they judged the questions and to leave different wrong answer choices unmarked. The results of this study indicated users of Nedelsky's method should consider a two stage judgment procedure. (Author/DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED218361

RR-82-6

RESEARCH REPORT

A STUDY OF THE RELIABILITY OF NEDELSKY'S METHOD FOR CHOOSING A PASSING SCORE

Samuel A. Livingston
William Kastrinos

March 1982

TM 820474

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Livingston, S.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Educational Testing Service
Princeton, New Jersey

A Study of the Reliability of Nedelsky's Method for Choosing a Passing Score

Samuel A. Livingston
William Kastrinos

Center for Occupational and Professional Assessment

Educational Testing Service
Princeton, New Jersey
March 1982

Copyright © 1982. Educational Testing Service. All rights reserved.

ABSTRACT

Licensing and certification boards have begun to use Nedelsky's method of choosing passing scores. Alternate forms of a test, containing 25 questions in common, were used to investigate the reliability of passing scores chosen by this method. Seven content experts judged both forms of the test. The differences between these judges in their passing scores for the 25 common questions were large and consistent. The judges tended to set higher passing scores the second time they judged the questions, and to leave different wrong answer choices unmarked. These results suggest that users of Nedelsky's method should consider a two-stage judgment procedure.

A Study of the Reliability of Nedelsky's Method for Choosing a Passing Score

Samuel A. Livingston
William Kastrinos

In 1954, Leo Nedelsky suggested a method for determining "absolute grading standards for objective tests" -- specifically, for multiple-choice tests. Nedelsky's method calls for a group of judges to examine each test question and "cross out those responses which the lowest D-student should be able to reject as incorrect." These judgments are used to determine the probability that such a test-taker would answer the question correctly, under the assumption that the test-taker would guess at random from those choices not rejected as incorrect. These correct-answer probabilities are then used in computing the expected test score for this hypothetical test-taker. Nedelsky suggested that the passing score be chosen to give this "F-D student" some specified probability of passing the test.

In recent years, Nedelsky's method has become increasingly popular for choosing passing scores, not only on academic achievement tests, but also on professional certification and licensing examinations. For most applications, the "F-D student" has become the "borderline test-taker" or the "minimally qualified test-taker." Typically, the expected score for this hypothetical person (as determined from the judgments of the individual questions) is chosen as the passing score.

How reliable is this method of choosing a passing score? This question is important, because for many of the test-takers the choice of a passing score will make the difference between passing and failing. If the passing score is chosen by an unreliable procedure, the validity of the resulting pass/fail decisions is questionable, at least for those test-takers whose

scores lie near the passing score. Anyone concerned about the reliability of the individual test-takers' scores should be equally concerned about the reliability of the standard with which those individual scores will be compared.

Reliability is not a single characteristic of a test or other measurement process. A test can have different reliabilities with respect to different sources of measurement error. The three sources of measurement error usually considered most important are the selection of specific questions or problems, the time of testing, and the scorer or grader. It is important that test scores be consistent over all three of these factors. Similarly, it is important that the judgments involved in choosing the passing score have the same three types of consistency. Several previous studies have investigated the inter-judge reliability of the judgments and/or the passing scores produced by Nedelsky's method, in the context of a comparison with one or more other methods (Andrew and Hecht, 1976; Brennan and Lockwood, 1980; Koffler, 1980; Rock, et al, 1980; Skakun and Kling, 1980; Archambault, et al, 1981; Baron, et al, 1981; Poggio, et al, 1981). However, none of these studies investigated the extent to which Nedelsky-type judgments are stable over time--even very short intervals of time. The purpose of the present study was to investigate the short-term stability of the judgments and the passing scores produced by Nedelsky's method, while controlling for other sources of unreliability.

Method

An investigation of short-term stability generally involves repeated measures--in this case, two rounds of judgments of the same test questions by the same people. The data for this study were collected as part of an actual application of Nedelsky's method to a licensing examination for real estate brokers. The purpose of the judging session was not to provide data for this study, but to choose passing scores for the test. The method was applied to two forms of the examination. The two forms will be referred to in this report as "Form X" and "Form Y." Each form consisted of 80 four-choice questions. Twenty-five questions appeared on both forms of the test, in the same sequential positions on both forms. (For example, question #34 on Form X was exactly the same as question #34 on Form Y.) The results presented in this report refer only to the 25 questions that appeared on both test forms.

The judgments required for Nedelsky's method were provided by seven persons representing the real estate commissions of seven different states. Two were legal counsel to their commissions, one was a director, one was an education director, two were chairmen, and one was simply a member. These "judges" judged one form immediately after the other. Four of the judges ("Panel A") judged Form X first; the other three ("Panel B") judged Form Y first. The judges were not told that the two forms of the test would have any questions in common.

Results

Each of the 25 questions contained one correct answer and three wrong answers. The judge's task was to make a judgment about each wrong answer. Therefore, the 25 questions required 75 separate yes-or-no judgments from each judge--the same 75 judgments on each of the two rounds of judging.

Table 1 and Figure 1 show the percentage of the wrong answers that each judge marked (to indicate that a minimally competent test-taker would recognize the answer as wrong) on each round of judging. Several facts are apparent from these data. First, the individual judges differed greatly in the percentage of wrong answers they marked. Even when the comparisons are only between judges on the same panel, to control for other factors, the differences are large (and statistically significant: $F = 8.38$ with 5 and 4 df; $p < .05$). Second, regardless of which form was judged first, the percentage of wrong answers marked tended to increase substantially from the first judging to the second judging. (This effect is statistically significant: $F = 12.27$ with 1 and 4 df; $p < .05$.) Seven of the eight judges marked more wrong answers on the second judging than on the first. The increase was greater for the judges on Panel B than for those on Panel A. (This interaction effect is statistically significant: $F = 36.66$ with 1 and 4 df; $p < .01$.) There was a slight tendency for the judges to mark more wrong answers to the 25 common items when they appeared on Form X than on Form Y, (81 percent versus 78 percent), but this effect is not statistically significant ($F = 2.93$ with 1 and 4 df).

Overall, 83 percent of the judgments of individual wrong answers were consistent from the first judging to the second. Some of the judges were much more consistent than others. The two judges who marked nearly all

Table 1. Percentages of wrong answers marked for elimination, for 25 questions appearing on both test forms.

	1st Time	2nd Time	Both Times	Neither Time
	(Form X)	(Form Y)		
Judge A1	96	95	95	04
A2	77	82	77	17
A3	97	100	97	00
A4	62	70	51	20
	(Form Y)	(Form X)		
B1	72	87	72	13
B2	64	79	56	13
B3	66	68	51	17
Panel A	83	87	80	10
B	68	78	60	15
All judges	76	83	71	12

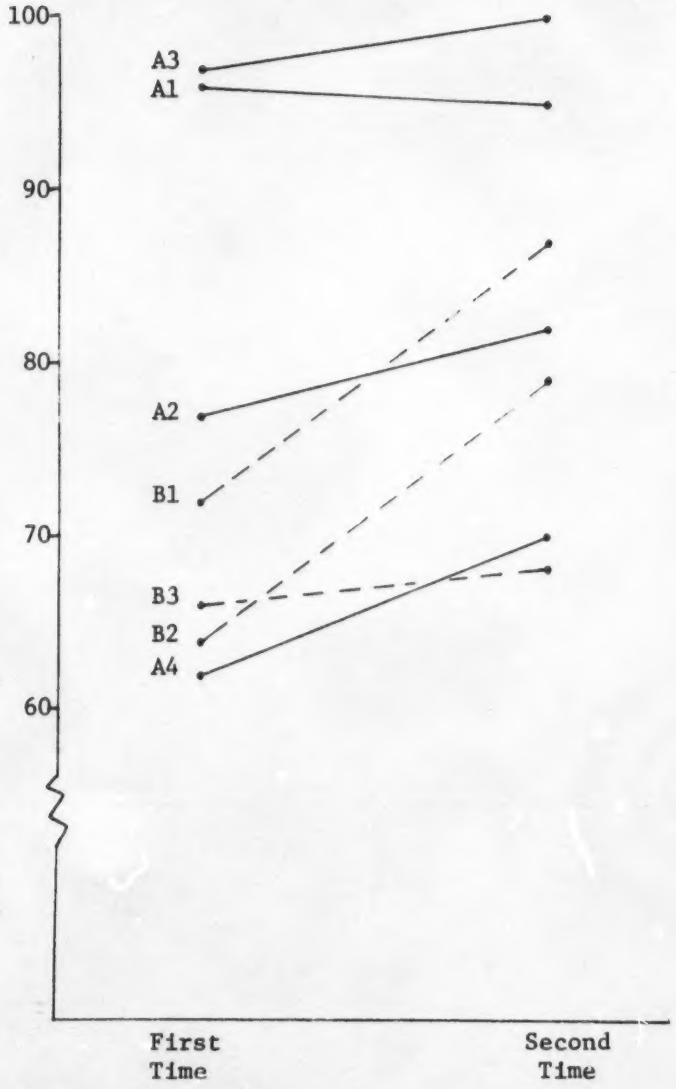

Figure 1. Percentages of wrong answers marked by individual judges.

of the wrong answers (Judges A1 and A3) were highly consistent. The two judges who marked the next largest numbers of wrong answers (Judges A2 and B1) were highly consistent in continuing to mark the answers that they had marked the first time. However, they also marked several answers the second time that they had not marked the first time. The remaining three judges (Judges A4, B2, and B3) made several changes of both types, though they tended to mark more answers the second time than the first. That is, on seeing a question for the second time, they were likely to think that the minimally competent test taker would eliminate more wrong answers than they had thought the first time they saw the question. Averaging over the seven judges, the average size of the differences between the percentages of wrong answers marked by the same judge on the two rounds of judging was 7.0 percentage points. As a basis for comparison, the average size of the differences between the percentages marked by different judges judging the same test form on the same round was 15.7 percentage points.

Table 2 and Figure 2 show the passing scores determined from the judgments by each judge in each of the two rounds of judging. These passing scores are simply the expected scores for the minimally competent test-taker, as computed from the judgments made by each individual judge. These results are, of course, highly similar to those for the number of wrong answers marked. They are not identical, because two answers marked in the same question have less effect on the passing score than do two answers marked in two different questions. The passing scores vary greatly between individual judges, especially on Panel A, where they range from 14.25 to 23.50 for the first round of judgments and from 15.83 to a perfect 25.00 for the second round. (This effect is statistically significant: $F = 14.44$ with 1 and 4 df; $p < .05$.) Seven of

Table 2. Passing scores determined from wrong answers marked, for 25 questions appearing on both test forms.

		1st Time	2nd Time
		(Form X)	(Form Y)
Judge	A1	23.50	23.00
	A2	16.83	18.50
	A3	24.00	25.00
	A4	14.25	15.83
		(Form Y)	(Form X)
	B1	18.92	21.83
	B2	16.42	20.00
	B3	17.17	18.08
		(Form X)	(Form Y)
Panel A:	mean	19.65	20.46
	S.D.	1.28	4.00
		(Form Y)	(Form X)
Panel B:	mean	17.50	19.97
	S.D.	1.28	1.88
All judges:	mean	18.73	20.25
	S.D.	3.70	3.04

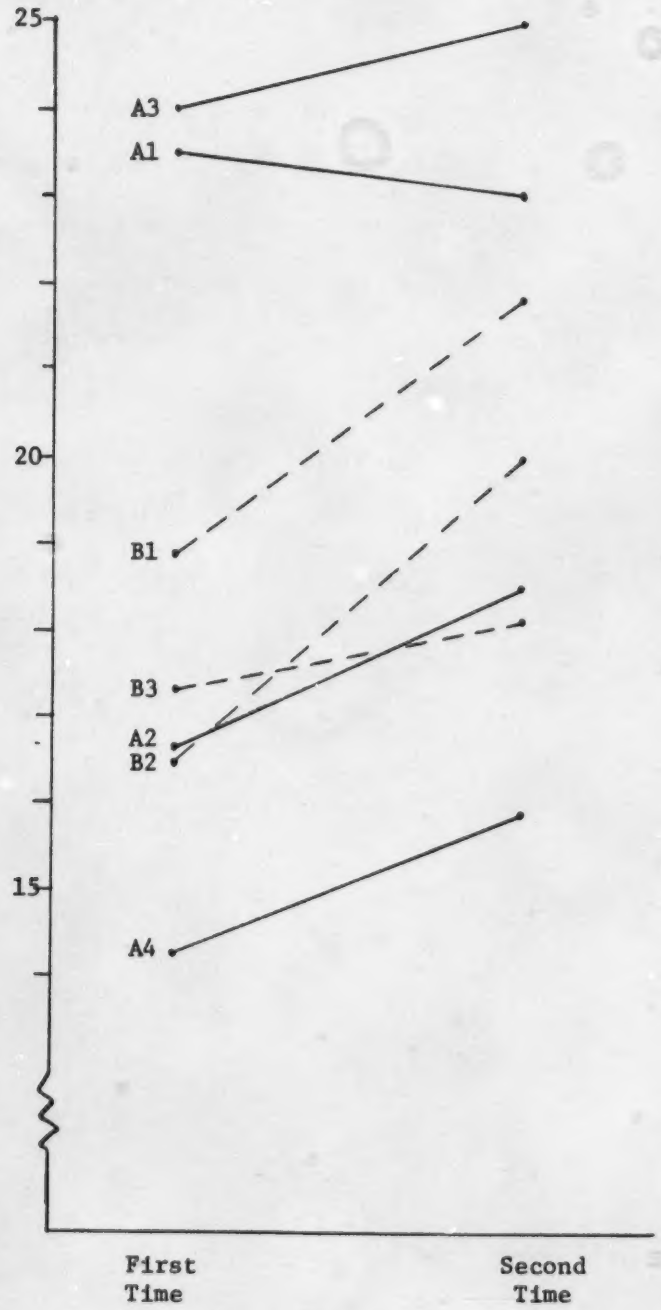

Figure 2. Passing scores determined by individual judges.

the eight judges set a higher passing score the second time than the first. The increase was somewhat greater for the judges on Panel B, who showed an average increase of 2.47 score points as compared with 0.81 for Panel A. (This effect is statistically significant: $F = 9.42$ with 1 and 4 df; $p < .05$.) The difference between the two different forms in the passing scores set for the 25 items was only 0.83 score points and is not statistically significant ($F = 2.91$ with 1 and 4 df). The average size of the differences between the passing scores set by the same judge on the two rounds, averaging over the seven judges, is 1.74 score points. As a basis for comparison, the average size of the differences between the passing scores set by different judges judging the same test form on the same round was 4.47 score points.

Discussion

The results of this small-scale study indicate that it is difficult to generalize about the consistency of the judgments collected and the passing scores determined by Nedelsky's method. Just as individual judges differ in their standards for acceptable performance, they differ in the consistency with which they apply those standards. Of course, a very high standard effectively guarantees a high degree of consistency. A judge who marks nearly all of the wrong answers each time cannot avoid marking most of them both times. How consistent do the judgments appear when this effect is removed? The "kappa" statistic (Cohen, 1960) provides an answer to this question. It compares the observed proportion of consistent judgments with the proportion expected by chance, assuming that the judge was predisposed to mark exactly the number of answers that he or she did mark. A value of zero indicates agreement at the chance level, under this assumption. Perfect agreement is indicated by a value of 1.00. The values of kappa for the seven judges in the study were -.08, -.40, .00, -.31, -.30, -.20, and -.23. These values indicate a tendency for the judges to leave different answers unmarked on the two rounds of judging. Although this tendency would not necessarily be reflected in the choice of a passing score, it may cast some doubt on the meaningfulness of the judgments.

The results of the study also show a clear tendency for judges to mark more answers the second time they judge the same questions. (Again, the judges vary considerably in the extent to which their judgments reflect this tendency.) This tendency makes the results of the study difficult to interpret, since the differences between the two rounds of judgments reflect this systematic trend

as well as the inconsistencies that the study was intended to investigate. Standardizing the data from each round of judgments is not an adequate solution, because of the obvious "ceiling" effect,* and because the absolute size of the differences is important.

What, then, are the implications of this study for potential users of Nedelsky's method? One possible implication is that users of the method should seriously consider using a two-stage procedure for collecting the judgments. In this procedure, the judges would judge all the test questions twice, then compare their judgments on the two rounds and resolve any discrepancies. This modification could be expected to result in judgments--and therefore, in passing scores--that are a better representation of the judges' standards. If time and circumstances permit, the judges might also be given the opportunity to discuss their judgments, before the final round. The opportunity for discussion might help to decrease the differences between individual judges, which is the largest source of unreliability in the judgments and in the resulting passing scores.

---

*That is, a judge cannot mark more than 100 percent of the wrong answers.

REFERENCES

Andrew, B. J. and Hecht, J. T. A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 1976, 36, 45-50.

Archambault, F. X., Behuniak, P., and Gable, R. K. A comparison of the Angoff and Nedelsky procedures for setting standards on reading and mathematics proficiency examinations. Paper presented to the National Council on Measurement in Education, Los Angeles, April, 1981.

Baron, J. B., Rindone, D. A., and Prowda, P. Setting statewide proficiency standards: A multi-method multi-fate analysis. Paper presented to the National Council on Measurement in Education, Los Angeles, April, 1981.

Brennan, R. L. and Lockwood, R. E. A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 1981, 4, 219-240.

Cohen, J. A coefficient of agreement for nominal sales. Educational and Psychological Measurement, 1960, 20, 37-46.

Koffler, S. L. A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 1980, 17, 167-178.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Poggio, J. P., Glasnapp, D. R. and Eros, D. S. An empirical investigation of the Angoff, Ebel, and Nedelsky standard setting methods. Paper presented to the American Educational Research Association, Los Angeles, April, 1981.

Rock, D. A., Davis, E. L., and Werts, C. An empirical comparison of judgmental standard-setting procedures. Research report RR-80-7. Princeton, N.J.: Educational Testing Service, 1980.

Skakun, E. N. and Kling, S. Comparability of methods for setting standards. Journal of Educational Measurement, 1980, 17, 229-235.